Explainable AI (XAI): Core Ideas, Techniques, and Solutions

Year: 2022 | Citations: 820 | Authors: Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, O. Rana

---

Abstract

As our dependence on intelligent machines continues to grow, so does the demand for more transparent and interpretable models. In addition, the ability to explain the model generally is now the gold standard for building trust and deployment of artificial intelligence systems in critical domains. Explainable artificial intelligence (XAI) aims to provide a suite of machine learning techniques that enable human users to understand, appropriately trust, and produce more explainable models. Selecting an appropriate approach for building an XAI-enabled application requires a clear understanding of the core ideas within XAI and the associated programming frameworks. We survey state-of-the-art programming techniques for XAI and present the different phases of XAI in a typical machine learning development process. We classify the various XAI approaches and, using this taxonomy, discuss the key differences among the existing XAI techniques. Furthermore, concrete examples are used to describe these techniques that are mapped to programming frameworks and software toolkits. It is the intention that this survey will help stakeholders in selecting the appropriate approaches, programming frameworks, and software toolkits by comparing them through the lens of the presented taxonomy.